Toward Predictive Modelling in Breeding of Tetraploid Potato

Sundmark, Ea Høegh Riis

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

TOWARD PREDICTIVE MODELLING IN BREEDING OF TETRAPLOID POTATO

BY
EA HØEGH RIIS SUNDMARK

DISSERTATION SUBMITTED 2019

AALBORG UNIVERSITY
DENMARK

TOWARD PREDICTIVE MODELLING IN BREEDING OF TETRAPLOID POTATO

BY
EA HØEGH RIIS SUNDMARK

AALBORG UNIVERSITY
DENMARK

DISSERTATION SUBMITTED 2019

Dissertation submitted: May 2019

PhD supervisor: Professor wsr Kåre Lehmann Nielsen
Aalborg University

PhD committee: Associate Professor Morten Simonsen Dueholm (chair.)
Aalborg University, Denmark

Head of Biotech and Turf Research Christian Sig Jensen
DLF Seeds A/S, Denmark

Senior Research Scientist Dan Milbourne
Teagasc, Ireland

PhD Series: Faculty of Engineering and Science, Aalborg University

Department: Department of Chemistry and Bioscience

ISSN (online): 2446-1636
ISBN (online): 978-87-7210-436-2

Published by:
Aalborg University Press
Langagervej 2
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2019

Preface

The present thesis is the result of an industrial PhD project carried out at Aalborg University, department of Chemistry and Bioscience in collaboration with Danespo A/S, KMC Amba and with funding from Innovation fund Denmark. The project was carried out in the period from April 2015 to March 2019. The aim of this industrial PhD project were to investigate the utility and implementation possibilities of molecular markers, in the form of haplotype detection in specific regions of the potato genome, in predictive modelling of phenotypic values for breeding purposes. To do this;

- A broad selection of molecular markers of different types (e.g. RFLP, SSR or SNP) and for different traits (yield, starch content, chip quality, tuber shape, maturity, *G. pallida* pathotype 2 resistance, *S. endobioticum* pathotype 1 and 6 and late blight resistance) from a literature study was converted into locus markers and anchored on the potato reference genome DM v4.03. Locus markers were used for haplotype detection in a diversity panel of 48 accessions of elite cultivars and breeding clones.
- Predictive models from two different approaches were tested for prediction accuracy and model fitness by utilizing detected haplotypes as predictor variables to predict phenotypic values of the diversity panel. For this part, historical data constituted the observed phenotypic values of the diversity panel.
- Locus markers were selected based on performance of markers in the predictive modelling of the diversity panel. These markers were applied to an offspring population of 94 accessions from a bi-parental cross between Lady Anna and breeding clone 04-EQF-6, to evaluate the effect of training population size when predicting different traits (yield, starch content, chip quality, tuber shape, maturity, *G. pallida* pathotype 2 resistance, *S. endobioticum* pathotype 1, 2 and 6 and late blight resistance).

Three scientific manuscripts and one popular science manuscript (manuscript 4) have been prepared, based on the findings:

Manuscript 1: “Anchoring molecular markers relevant for potato breeding onto the draft genome sequence of potato”.

At time of thesis submission, to be submitted to Potato Research - Journal of the European Association for Potato Research

Sundmark, E.H.R., Sverrisdóttir, E., Sønderkær, M., Kirk, H.G., Nielsen, K.L.

Manuscript 2: "Prediction of yield, starch content, maturity and late blight resistance in a diversity panel of breeding relevant potato".

At time of thesis submission, to be submitted to Theoretical and Applied Genetics - International Journal of Plant Breeding Research

Sundmark, E.H.R., Sverrisdóttir, E., Sønderkær, M., Lindskou, M., Kirk, H.G., Nielsen, K.L.

Manuscript 3: "Performance of predictive modelling of yield and starch content on a biparental cross of elite potato breeding germplasm".

At time of thesis submission, to be submitted to undetermined peer reviewed journal e.g. Molecular Breeding - New Strategies in Plant Improvement

Sundmark, E.H.R., Sverrisdóttir, E., Sønderkær, M., Kirk, H.G., Nielsen, K.L.

Manuscript 4: "When breeding new potato cultivars, haplotypes are the new black".

At time of thesis submission, to be submitted to undetermined popular science journal e.g. The Scientist – Exploring life, inspiring innovation

Sundmark, E.H.R, Sverrisdóttir, E., Sønderkær, M., Nielsen, K.L.

Furthermore, the following popular paper was produced as part of the project, but not included in the thesis: Sundmark, Ea Høegh Riis (2017) *Fart på forædlingen.* Dansk Kartoffelstivelse, Maj 2017, nr. 2

Acknowledgements

My sincerest thanks goes to my academic advisor, Kåre Lehmann Nielsen and to Hanne Grethe Kirk, who have taken up the role of industrial advisor. You have given me irreplaceable guidance and fruitful discussions throughout this journey.

I also want to thank Elsa Sverrisdóttir and Mads Lindskou, who have helped me with data processing and development of scripts. No matter how small you think your contributions were, they have been a tremendous help for me.

I would like to express my gratitude to all the colleagues I have gained through this collaboration between AAU, KMC and Danespo. All of you have welcomed me with interest and positive attitudes, even when you only met me once or twice a month. Thanks to Anne, for figuring out what to do, when my DNA extractions didn't work. Thanks to Mette and Elsa for discussions and laughs in the office, to Mads for his extensive work with troubleshooting of the data pipeline and to all the other members of the Functional Genomics group at the Department for Chemistry and Bioscience for creating a good work environment. Thanks to all of the people at Danespo R&D for helping me take care of and test my offspring population and for teaching me about potato breeding in practice. Especially thanks to Kirsten, for testing the chip quality though the tubers were too small or too large or rubbery.

Last but definitely not least, I would like to thank my wonderful family and friends. Without your support, encouragement and understanding I would surely have given up along the way. Thank you mom, dad, Maya, Kristen and Kaspar for your feedback on all the text I made you read and especially to Kaspar for putting up with me, even on my worst days. Thanks also to Kamilla and Camilla for always having time for a cup of tea and a talk.

All of you are invaluable to me

English summary

The food industry is currently experiencing a demand for higher-yielding food crops for a globally growing population. At the same time, higher and higher demands are being made for reducing pesticides, which in particular affects the potato industry, as it is dependent on pesticides to control late blight in the field, among other things. In the pursuit to increase the acceleration of genetic gain in the breeding of new potato cultivars to meet the requirements, molecular markers have been developed since the late 1980s to analyze phenotypic traits in the breeding germplasm. In recent years, the focus has been on methods utilizing high-throughput sequencing of DNA to examine large amounts of SNP markers at once. Genomic selection is one of these methods and it has been shown to successfully estimate a genomic breeding value (GEBV) from genome wide SNP markers through various Bayesian methods or Best Linear Unbiased Prediction (BLUP) in many populations and traits. However, it is unfortunately also very expensive to implement as it may need 100,000+ markers.

The overall purpose of this thesis has been to investigate a possible replacement for genomic selection in the form of predictive modelling based on detected haplotypes found in a genetic diversity population called the diversity panel and an offspring population, called F1 population, obtained from two parents from the diversity panel, respectively. By utilizing the fact that haplotypes gives higher resolution of the underlying allele structure in the regions of the tested locus markers, it is theoretically possible to reduce the number of markers needed. Each of these markers have higher linkage between marker and trait. Through the work presented in manuscripts 1 and 2 of this thesis, a set of locus markers is defined from a thorough literature study and these are applied to the diversity panel to detect haplotypes. Genotype information of the detected haplotypes is subsequently used to make predictive models from two different approaches. The first approach is based on multiple linear regression and the other on regression trees. The studies show that the utility of the two methods depends on the trait in question and the population size. High prediction accuracy for both approaches is demonstrated in all traits, but also a large prediction error due to overfitting. In manuscript 3, a proportion of the markers are applied to the F1 population and used to investigate the applicability of the methodology in a population of full siblings. Here, the accuracy of the models was lower than expected, which may be due to the composition of the marker assay. Furthermore, the thesis contains the results of the two approaches of predictive models applied to the diversity panel for the traits that have been studied through the collaboration with Danespo A / S, but which have not been adapted to publication. These results also show high prediction accuracies, but also a higher degree of overfitting, due to smaller

training population size because of missing phenotypic values of some individuals in the diversity panel.

Overall, the results of this thesis indicate that detected haplotypes can be used for predictive modelling to estimate the phenotype from a DNA sample with good results. Even for traits where it was only possible to obtain phenotypic values for less than 30 individuals of the diversity panel, a phenotypic value could still be estimated with high accuracy. However, this was only possible within the diversity panel. The best results were obtained when the training population was at least 34 individuals. These results indicate that it is indeed possible to make useful and robust models for the breeding of tetraploid potatoes based on haplotype detection with PCR amplification and high-throughput sequencing.

Danish summary

Fødevareindustrien oplever for tiden en efterspørgsel for højere ydende afgrøder til fødevarer til en globalt voksende befolkning. Samtidig bliver der stillet højere og højere krav til reduktion af sprøjtemidler, hvilket især påvirker kartoffelindustrien, der er afhængig af sprøjtemidler for at kontrollere blandt andet skimmel angreb i marken. I jagten på at forøge accelerationen af genetisk gevinst i forædlingen af nye kartoffelsorter, der kan imødekomme kravene, har man siden slutningen af 1980'erne udviklet molekylære markører til at undersøge fænotypiske træk hos planter i forædlingsprogrammerne. I de seneste år har man fokuseret på metoder, der udnytter high-throughput sekventering af DNA til at undersøge store mængder af SNP markører på én gang. Genomisk selektion er én af disse metoder, som er bevist at kunne estimere et genomisk avlsværdital (GEBV) ud fra SNP markører spredt over hele genomet gennem forskellige Bayesian metoder eller bedst mulig liniære objektive forudsigelse (BLUP) i forskellige populationer og for forskellige træk. Det er dog desværre også en meget dyr metode at implementere, da det kan være nødvendigt at bruge 100.000+ markører.

Det overordnede formal med denne afhandling har været at undersøge en mulig erstatning for genomisk selektion i form af prædiktiv modellering ud fra påviste haplotyper fundet i hhv. en genetisk divers population kaldet diversitets panelet og en afkomspopulation af to forældre fra diversitets panelet. Den sidstnævnt kaldes F1 populationen. Ved at udnytte at haplotyper bedre beskriver den underliggende allel struktur i de områder hvor de påvises, er det teoretisk muligt at nøjes med færre markører, der så hver især har højere kobling mellem markører og egenskab. Gennem arbejdet præsenteret i manuskript 1 og 2 af denne afhandling defineres et sæt af locus markører ud fra et gennemgribende litteraturstudie og disse anvendes på diversitets panelet for at påvise haplotyper. Genotype information om de påviste haplotyper anvendes efterfølgende til at lave prædiktive modeller ud fra to forskellige fremgangsmåder. Den første fremgangsmåde er baseret på multiple lineær regression og den anden på regressions træer og undersøgelserne viser at anvendeligheden af de to fremgangsmåder afhænger af hvilket træk der undersøges. Der påvises stor nøjagtighed af forudsigelserne for begge fremgangsmåder i alle træk, dog også en stor forudsigelsesfejl grundet overfitting. I manuskript 3 anvendes en andel af markørerne på F1 populationen for at undersøge anvendeligheden af metodikken i en population af helsøskende. Her findes nøjagtigheden af modellerne til at være lavere end forventet, hvilket muligvis kan skyldes sammensætningen af markører. Yderligere indeholder afhandlingen resultater af de to fremgangsmåder af prædiktive modeller anvendt på diversitetspanelet for de træk som er blevet undersøgt igennem samarbejdet med Danespo A/S, men som ikke har kunnet tilpasses publicering. Også disse

resultater viser høj nøjagtighed af forudsigelserne, men også en højere grad af overfitting, hvilket skyldes mindre træningspopulations størrelse som følge af manglende værdier af fænotypen for nogle individer i diversitets panelet.

Generelt er resultaterne i denne afhandling et udtryk for at påviste haplotyper kan anvendes i prædiktive modeller til at estimere fænotypen ud fra en DNA prøve med gode resultater. Selv for træk hvor det kun var muligt at tilvejebringe fænotype værdier for få individer i diversitets panelet kunne der stadig estimeres en værdi af fænotypen, dog kun inden for diversitets panelet og de bedste resultater blev opnået når træningspopulationen var på mindst 34 individer. Disse resultater indikerer at det er muligt at lave anvendelige og robuste modeller til forædlingen af kartofler baseret på påvisning af haplotyper med PCR amplificering og high-throughput sekventering.

Table of contents

Introduction

With a growing global population, the world's food supply needs to grow with it (Valin et al. 2013). Potato is one of the most promising food crops to meet this need with a potential for high calorie per hectar compared to other crops (Horton 1980). To develop new elite cultivars, that can fulfil this potential, breeders need tools to improve their selection methods and increase the rate of genetic gain. To this extent, molecular markers has been proven effective in breeding of many crops including potato (Xu and Crouch 2008, Tiwari et al. 2013).

The history of molecular markers in potato breeding

In the general context of plant breeding, molecular markers was first used in the early 1980s when isozyme markers were introduced and a little later with the first use of restriction fragment length polymorphism (RFLP) (Botstein et al. 1980) markers in 1986. These molecular markers was used in potato breeding shortly after that, with diploid linkage maps based on RFLP markers from tomato in 1988 and potato RFLP markers in 1989 (Barrell et al, 2013). This type of marker dominated throughout most of the 1990s, foremost as anchor points in map comparisons. Despite being able to capture genetic changes and predict linkage relationships between loci of different species, RFLP markers are cumbersome to work with and therefore limiting the amount of plants that can be screened at a time (Gebhardt et al. 1989, Xu and Crouch 2008). In comparison, the PCR-based amplified fragment length polymorphism (AFLP) (Vos et al. 1995) markers is easier to use and can screen up to 50 specific loci at a time (Rouppe van der Voort et al. 1998, Poczai et al. 2013). Hence, AFLP markers followed RFLP markers in mapping studies and was still used in 2008 where D'hoop et al. (2008) used this type of marker for an association study. However, the AFLP markers are anonymous and mostly dominant, and therefore unfit for population structure analysis and cannot be readily anchored to the recent potato genome sequence (PGSC 2011). Also during the 1990s, randomly amplified polymorphic DNA (RAPD) (Williams et al, 1990) markers was developed and among other uses, Jacobs et al (1996) used RAPD markers for fine mapping of major genes for traits of interest in potato. Simple sequence repeat (SSR) markers (Kit 1961) (also called microsatellites) soon superseded RAPD markers due to better reproducibility and reliability. SSR markers was first reported in potato in 1996 and has since been used for identifying SSR locations on potato linkage groups and constructing linkage maps for quantitative trait loci (QTL) mapping in potato as well as fingerprinting and identifying potato

germplasm accessions and cultivars (Hirsch et al. 2016, Barrell et al. 2013). Recently, Reid et al. (2011) showed the usefulness of SSR markers for profiling European potato varieties and the INRA BrACySol Biological Resource Center (UMR IGEPP, Ploudaniel, France) use SRR markers for genotyping their current potato cultivar collection (Esnault et al. 2016). Another method for fingerprinting of potato germplasm is the diversity array technology (DArT) (Jaccoud et al. 2001), first developed for the rice genome and later used in potato by Śliwka et al. (2012). The sequence-characterized amplified region (SCAR) (Paran and Mickelmore 1992) marker system followed the RAPD, SSR, RFLP and AFLP markers. This system can identify a unique locus because it is based on sequenced RAPD PCR products for specific oligonucleotide primers. It was originally developed for lettuce, but have been reported in potatoes by Jansen van Rensburg and Dubery (2001). The cleaved amplified polymorphic sequence (CAPS) (Konieczny and Ausubel 1993) markers are based on specific oligonucleotide primers as well, but also utilize at least 25 restriction enzymes to target restriction site polymorphisms in the PCR product to distinguish different genotypes. CAPS markers were originally developed for *Arabidopsis thaliana*, but De Jong et al. (1997) converted AFLP markers into CAPS markers for use in mapping of a hypersensitive potato virus X resistance gene and more recently, Sulli et al. (2017) used CAPS markers for molecular characterization of a panel of tetraploid and diploid potato genotypes.

Molecular markers in selection methods

As the market has grown to have higher demands for multiple disease resistances together with high yields and quality traits in cultivars, selection has become more comprehensive for breeders. Breeders traditionally select candidates for sexual parental crossings based on phenotypic performances. More recently, selection methods, such as Marker Assisted Selection (MAS) (Milczarek et al. 2011), has been applied to analyze for specific genes known to give resistance against diseases. With this it is possible to specifically identify which disease resistant cultivars complements others and in that way pyramid disease resistance genes in a single cultivar (Tomczyńska et al. 2014). The offspring with presence of the molecular markers are then selected for further breeding. MAS in potato, has focused on the large, often dominant, contributions from a limited number of molecular markers (Slater et al. 2016). Each marker is evaluated for presence or absence independently of other markers, as linkage between markers are often not studied. In potato breeding, selection for multiple traits with MAS is a costly process, as it is

necessary to produce a large number of offspring to ensure high likelihood of combined traits in the offspring population. The analysis of marker presence in each offspring is often a cumbersome and time-consuming task.

In contrast to MAS, Genomic Selection (GS) (Meuwissen et al. 2001) strives to predict phenotypes based on small contributions from genome-wide markers (Heslot et al. 2015). It also has the benefit of only requiring one DNA sample from each individual, which then are tested for multiple markers, e.g. with SNP chip (Wang et al. 1998). Though not commonly implemented in potato breeding, different studies show good prospect of GS in predicting phenotypes in-population (Enciso-Rodriguez et al. 2018, Arruda et al. 2015, Heslot et al. 2015). However, a study done by Sverrisdóttir et al. (2018) showed a large drop in correlations of the prediction models when predicting out-of-population and Stich and Melchinger (2009) suggested that the prediction accuracies obtained in their study were due to modelling of relatedness more than due to linkage between markers and quantitative trait locus (QTL). GS often uses statistical methods such as best linear unbiased prediction (BLUP), Bayesian generalized linear regression (Bayes) or reproducing kernel Hilbert spaces (RKHS) (Morrell et al. 2012, Habyarimana et al. 2017). All of these are dependent on Single Nucleotide Polymorphisms (SNPs) and developed for genomic analysis of far less diverse genomes as the human genome or Arabidopsis thaliana, which have 1 SNP per 1300 base pairs (bp) and 1 SNP per 3.300 bp, respectively (HGSC 2001, Kaul 2000). When considering the SNP density in the potato reference genome of 1 SNP per 40 bp (PGSC 2011), it becomes more apparent why correlations drop when predicting phenotypes in an out-of-population manner in potato. Even when using over 170.000 SNP's to calculate genomic estimated breeding values as done by Sverrisdottír and Byrne et al. (2017), it is unlikely that many of those SNPs are distinguishing only two alleles (manuscript 1, D'hoop et al. 2014, Esnault et al. 2016). Consequently, the SNPs will be in varying degree of association to traits dependent on which group of alleles the SNP represents and the frequency of the causal allele within that group. Linkage disequilibrium to QTLs are therefore not always faithfully estimated in the calculations. To counteract the loss of linkage, because of SNPs associating to multiple alleles, sequence variation in a region of interest can be combined into haplotype markers to obtain higher resolution of the allele structure.

Barriers in implementation

Though genetic markers have been used, and are still used, in many scientific studies, the implementation into practical potato breeding have been limited (Ramakrishnan et al. 2015). One barrier for implementing molecular markers in practical breeding, is the very large number of different molecular biology techniques each breeding station will have to operate to utilize the different types of molecular markers. Another barrier is the high diversity, not only in potato as a species, but also between different breeding populations (Sverrisdóttir et al. 2018). Because of this, a marker with good associations to a specific trait in one population may not be diagnostic in other populations and simply implementing markers found by others in MAS often give limited results. Indeed, markers, whose effect was documented in diverse populations is more likely to give robust results in practical breeding. However, many existing markers have been identified in diploid germplasm and bi-parental segregating populations (Moloney et al., 2010, Stich et al., 2013), which are by nature less diverse. It is therefore necessary to test a range of different markers, of which expectedly only some will prove applicable for the breeding population and trait in question. Hence, a methodology, which considers the most suitable of the tested markers for use in predictive modelling, have been applied in our work. Furthermore, we utilize the higher resolution of haplotype markers amplified from specific loci on the potato reference genome DM v4.03 (PGSC 2011) to discover allele structure in sites of interest.

Statistical methods

By using locus specific markers (manuscript 1 and 2) to obtain haplotypes for use in Genomic Prediction models, we can obtain more robust and unique 1:1 linkage between marker and trait and predict multiple traits from the same set of widespread molecular markers. For this purpose, we have studied two approaches for Genomic Prediction, a Supervised Genomic Prediction (SGP) approach and a Regression Tree (RT) approach. For the SGP approach, a combination of quality filtering for noise haplotypes, subset selection of significant haplotypes and a following Multiple Linear Regression (MLR) model form the basis of our proposed method. Gene effects are assumed to be additive, which makes it possible to create a very simple stepwise addition of constants chosen based on which variable gives the greatest additional improvement to the model fit. The Akaike information criterion (AIC) for each fit is used to evaluate which model gives the best fit. Starting with a null model containing only an intercept $Y = \beta_0$, the stepwise

selection includes more haplotypes as predictor variables (X) in the model step by step and the model takes the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

for p different haplotypes as variables. β is the slope coefficient for the predictor variable against the response and ϵ is the error term (James et al. 2013). MLR models assumes a linear relationship between response and predictors, and if this is true, the model has low bias. However, unless $n \gg p$, MLR models tend to have high variance and are prone to overfitting. As phenotypic values (n) are more costly and time-comsuming to produce than genotypic observations (p), overfitting can be counteracted by constraining predictor variables to only those of highest relevance for each model. In the SGP approach, this is implemented in two steps of quality filtering and subset selection, respectively. Quality filtering with offset in the Pearson correlation coefficient between observed phenotype and genotype evaluates the strength of the relationship between a given haplotype and the observed phenotype. A Pearson correlation coefficient of less than $|0.2|$ can be considered negligible and therefore regarded as noise (Mukaka 2012). While relationships between variables can lead to an inflated significance of predictor variables, the subsequent use of stepwise selection for the MLR models adjusts the significance of each predictor for the other predictors and thereby only selects the most important predictor variables from the available subset. The Root Mean Squared Error (RMSE) is used to evaluate the robustness of the resulting multiple linear models (James et al. 2013).

Decision Trees is the collective name for Classification and Regression Trees, and is a statistical method of selecting important predictor variables from a large number of detected variables for determining the response variable (Breiman 1984). A tree structure is grown from the root node by segmenting response data into two distinct groups (e.g group A and B) by the predictor variable that best describes the difference between the groups (p_1). This constitutes the internal nodes. These groups are then segmented once again into exclusive groups using the next predictor variable for each of group A and B. Hence, group A is split with p_2 and group B with p_3 and this is repeated until the number of observations in the resulting nodes is two. This recursive binary partition selects the predictor variables that minimize the residual sum of squares, the difference between the observed response and the response predicted by the model. In the resulting nodes, the leaf nodes, the mean of the resulting groups are given as the predictive phenotypic

value of all accessions in the node. Cost complexity pruning is applied to reduce overfitting by obtaining a set of best subtrees through cross-validation, which is in turn used to evaluate which pruning of the tree minimizes average error. To evaluate the robustness of each model a Cross-Validated Error Rate (CVER) can be estimated from the deviance between accessions in the leaf nodes. Decision Trees are generally considered more robust than linear models when the relationship between measured variables and observed responses are complex (James et al. 2013). Using this approach therefore has the possibility of including factors such as epistasis in the model, a factor that is in most cases not quantified in potato (Li et al. 2010, Rejwan et al. 1999).

Prediction and inference in modelling

When performing analysis in a setting where input in the form of predictor variable information is more readily available than the output in the form of response, it is possible to predict the response based on predictor variables. In such situations, the goal is to predict the response with high certainty and the components of the estimated function of underlying predictive relationship are often treated as a black box. The composition of this black box is less relevant as long as the predictions are accurate (James et al. 2013, Hastie et al. 2009). If the goal was to infer the separate effect of each predictor variable on the response, the true function of underlying predictive relationship is estimated through estimation of parameters such as β from the MLR model described above (James et al. 2013). From the same reasoning, correlation between predictor variables have less impact on the usefulness of a predictive model, as the effect of each predictor variable on the response does not need to be separated from the effects of other predictor variables. When predicting a response by models with a large number of variables X_p, such as MLR, every variable can be described by the remaining variables in the model and multi-collinearity becomes unavoidable. At that point, inference in the form of untangling the effects of each marker is not possible in a single data set, and the obtained model becomes one of many possible models, until it can be validated in another, independent data set (James et al. 2013).

Additional Results

Additional results, that have been produced throughout this PhD, but not included in the former manuscripts, are presented in this chapter. The traits presented in the additional results are chip quality, tuber shape, resistance against white potato cyst nematode *Globodera* pallida pathogen 2 and resistance against *Synchytrium* endobioticum pathogen 1 and 6, a fungus causing potato wart disease. All of the additional traits are scored on a scale from 1 to 9, with 1 being the dark chip, round tubers or full susceptibility and 9 being no darkening of chip, long tubers or full resistance.

Use of historical data as predictor variables in models

With the exception of tuber shape, all additional traits have a lower amount of phenotypic values, than those examined in manuscript 2 (yield, starch content, maturity and late blight resistance). Resistance to wart disease pathogen 6 has the lowest amount of measurements of phenotype with only 16 accessions having a recorded phenotypic value. Chip quality and resistance to *G. pallida* pathogen 2 both have recorded phenotypic values for 20 of 48 accessions in the diversity panel, resistance to wart disease pathogen 1 has recorded phenotypic values for 25 accesions and 35 out of 48 accessions in the diversity panel have a recorded tuber shape value. The distribution of phenotypic values of these traits can be seen in figure 1. Chip quality and tuber shape is approximately normally distributed around the mean (4.2 and 5.3, respectively), while there is a seperation into two distinct groups for disease resistance traits *G. pallida* resistance (figure 1C) and wart disease resistance (figure 1D and E). The normal distribution of chip quality and tuber shape is a result of the diversity panel consisting of elite cultivars and breeding clones for different market segments spanning from fresh market to processing. The distribution of resistance trait phenotypes reflects the presence of cultivars for different geographical regions, some of which requires specific disease resistance and some where it is not in the same way a requirement. It could also be a result of the diversity panel consisting of old and new cultivars, as some old cultivars may not have resistance genes against the current relevant diseases.

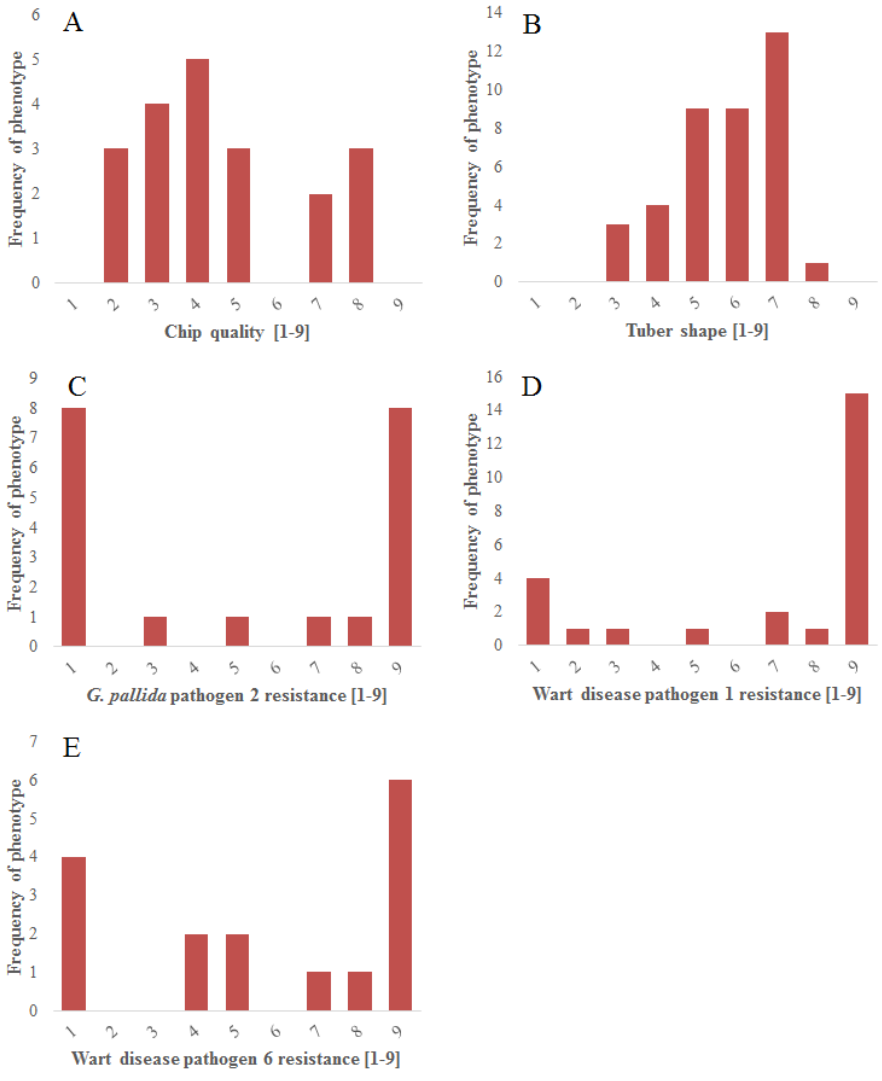

Figure 1: Frequency of phenotypic values for chip quality (A), tuber shape (B), resistance to G. pallida pathogen 2 (C), resistance to wart disease pathogen 1(D), and resistance to wart disease pathogen 6 (E).

Prediction of phenotypic values with Supervised Genomic Prediction models and Regression Tree models

All predicted phenotypic values in the following figures 2-7 are found with Leave-One-Out cross-validations of models obtained with the SGP approach and RT approach described in the introduction. For further details on the methods of model calculations, see manuscript 2. Models obtained with the SGP approach is given in figures 2-4. Coefficient of determination range from 0.51 for resistance to wart disease pathogen 6 (figure 4E, RMSE: 2.28) to 0.6 for chip quality (figure 2A, RMSE: 1.23) and 0.77 for both tuber shape (figure 2B, RMSE: 0.59), resistance to *G. pallida* pathogen 2 (figure 3C, RMSE: 1.79) and wart disease pathogen 1 resistance (figure 3D, RMSE: 1.52). Models from the RT approach are given in figures 5-7. Except for wart disease pathogen 1 resistance, the RT models give higher coefficient of determination than the SGP models. For the RT model of chip quality the coefficient of determination is 0.92 (figure 5A, CVER: 2.44), 0.87 for the tuber shape model (figure 5B, CVER: 1.13), 0.98 for the *G. pallida* pathogen 2 resistance model (figure 6C, CVER: 5.51), 0.41 for the wart disease pathogen 1 resistance model (figure 6D, CVER: 4.03) and 0.94 for the model of resistance to wart disease pathogen 6 (figure 7E, CVER: 4.02). Though the Coefficient of determination was higher for RT models, the RMSE values for the SGP model were lower than the CVER values of the RT model for every trait, suggesting more robust models are obtained with the GSP approach, as both evaluation parameters RMSE and CVER are relative to the observed response scale of 1 to 9. Sverrisdóttir et al. (2017) has reported chip quality heritability to be between 0.65 and 0.78 dependent on how the heritability was estimated. The prediction accuracies of 0.6-0.92 obtained with the models in this study enclose the range repoted by Sverrisdóttir et al. (2017), though the RT model is most likely overfitted, as reflected by the CVER value. Prediction accuracies of 0.77-0.87 for the tuber shape models are very close to the reported heritability of the trait of 0.8 (van Eck et al. 1994) and within the range of heritability reported by Willemsen (2014), who found heritabilities of 0.75-0.9 dependent on population and field conditions. Heritability of *G. pallida* pathogen 2 resistance is expected to range from 0.88 to 0.93 (Kreike et al. 1994, Finkers-Tomczak et al. 2014) and as we saw with chip quality, the the SGP model falls a little short of the expected value, while the R^2 of 0.98 for RT model is a little higher. Heritability of resistance to wart disease pathogen 1 is established by Groth et al. (2013) to be 0.85 while it has not been possible to find heritability values on wart disease pathogen 6, though Groth et al. (2013) estimates the cumulative R^2 of multiple regression based on effects from 4 molecular markers to be 0.31. Hence, the SGP models estimate the relationship of two resistance traits fairly well (0.77 and 0.51, respectively), while the RT models estimate the relationship to be opposite, though with high prediction errors of approximately 4.

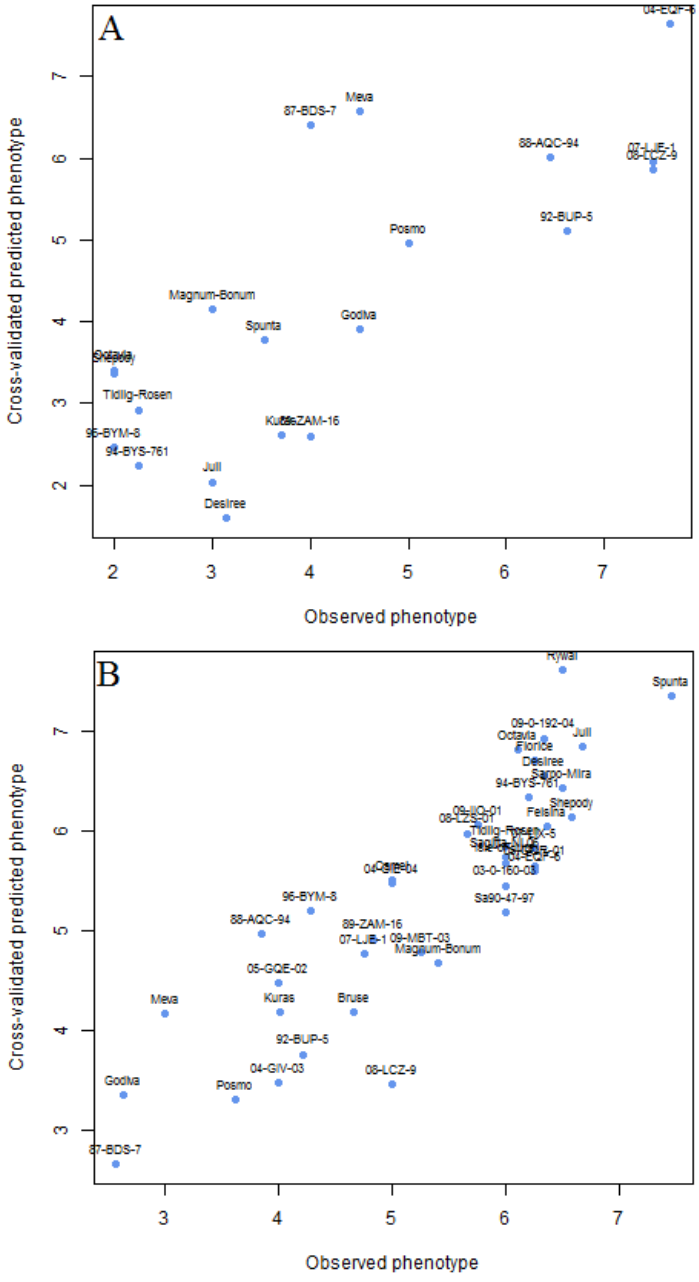

Figure 2: Supervised Genomic Prediction models of chip quality (A) (R^2=0.6, RMSE=1.23) and tuber shape (B) (R^2=0.77, RMSE=0.59).

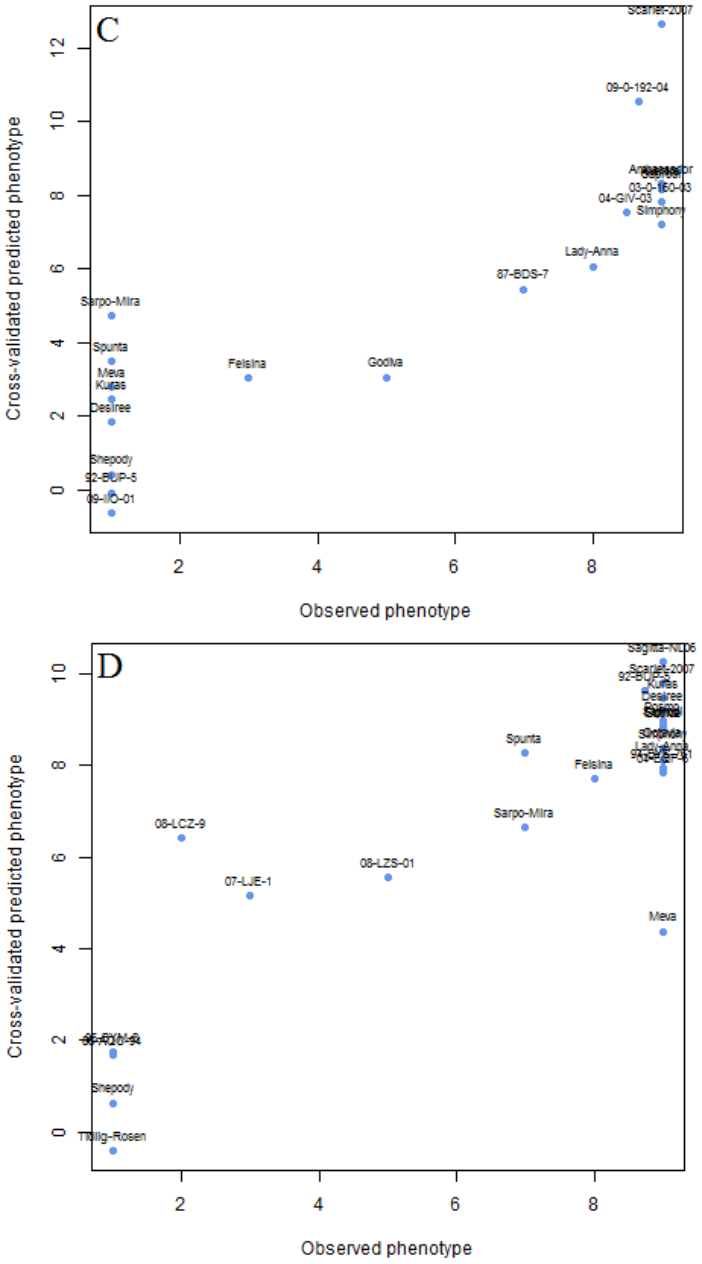

Figure 3: Supervised Genomic Prediction models of G. pallida Pa2 resistance (C) (R^2=0.77, RMSE=1.79) and wart disease pathogen 1 resistance (D) (R^2=0.77, RMSE=1.52).

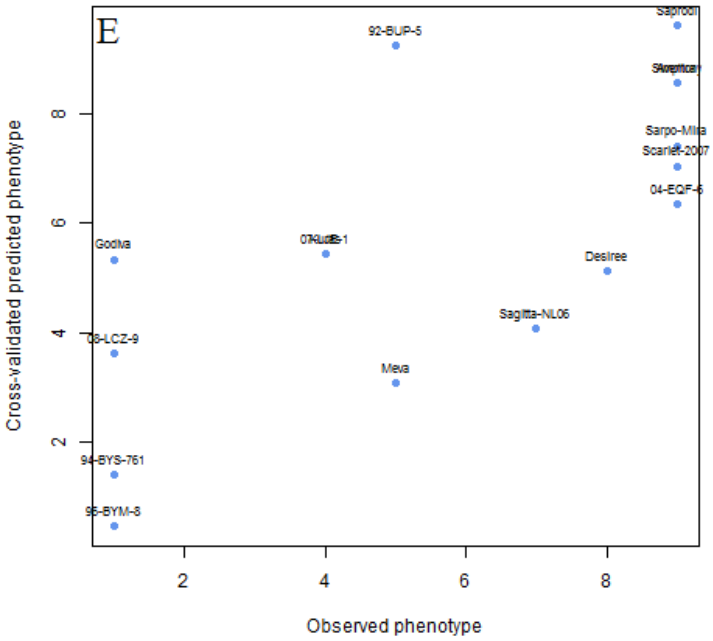

Figure 4: Supervised Genomic Prediction models of wart disease pathogen 6 resistance (E) (R^2=0.51, RMSE=2.28).

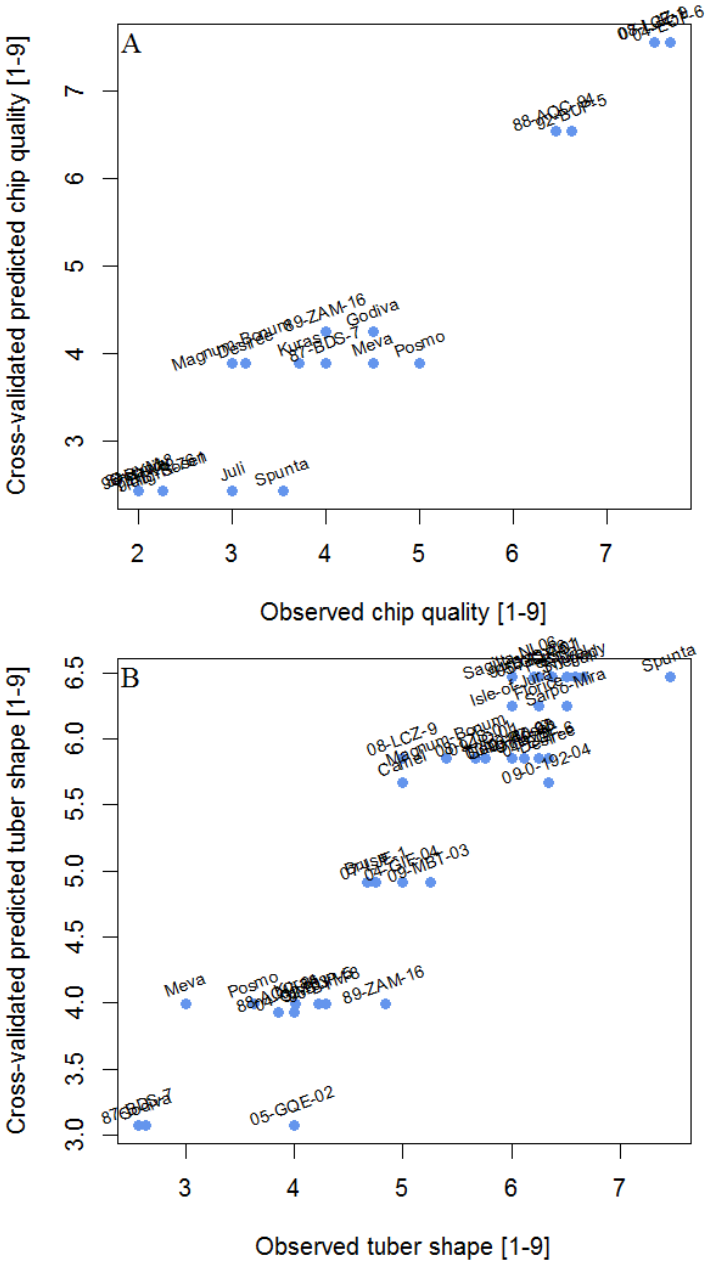

Figure 5: Regression tree models of chip quality (A) (R^2=0.92, CVER=2.44) and tuber shape (B) (R^2=0.87, CVER=1.13).

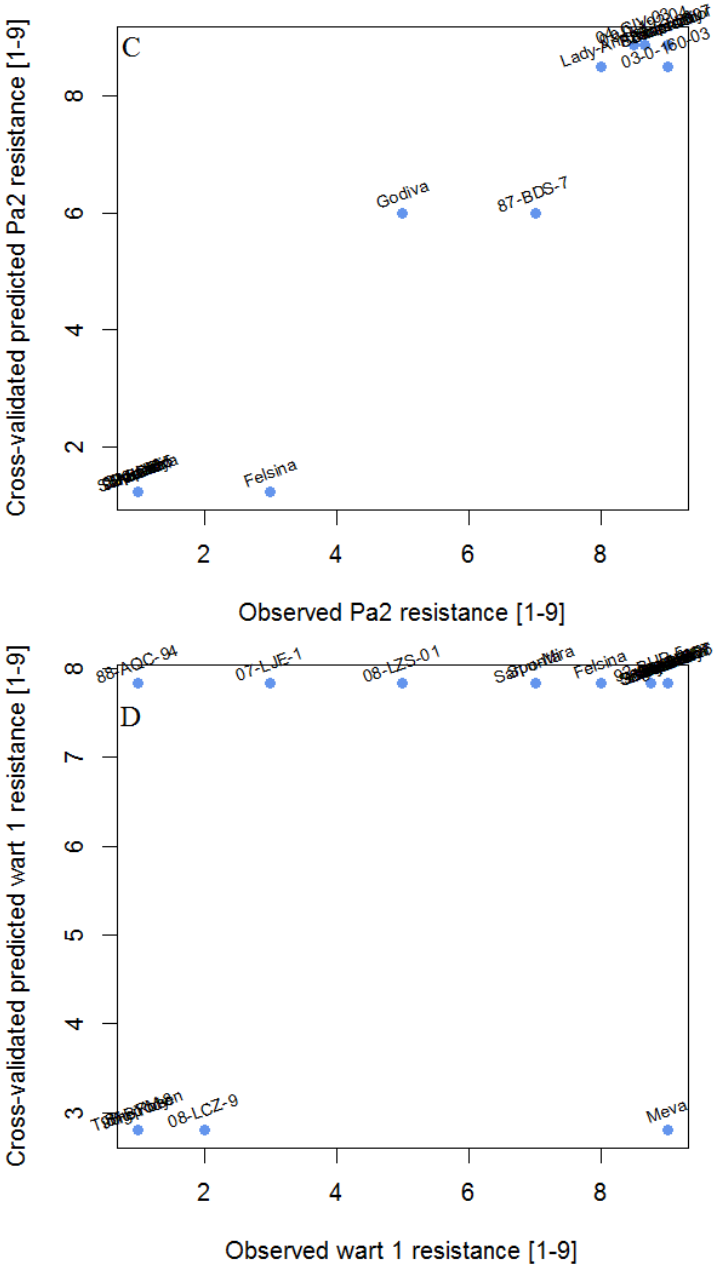

Figure 6: Regression tree models of G. pallida Pa2 resistance (C) (R^2=0.98, CVER=5.51) and wart disease pathogen 1 resistance (D) (R^2=0.41, CVER=4.03).

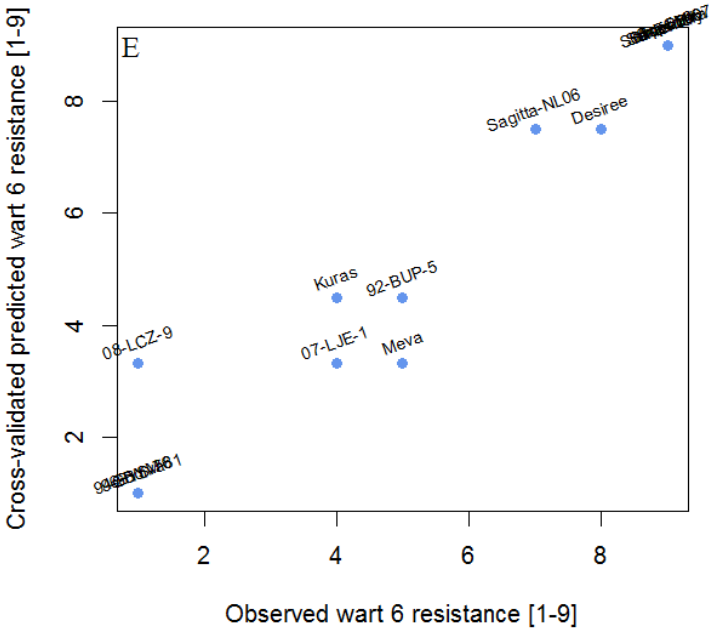

Figure 7: Regression tree model of wart disease pathogen 6 resistance (R^2=0.94, CVER=4.02).

General discussion

Utility of molecular markers of known marker-trait associations in potato

The basis of this thesis is the vast amount of knowledge already available regarding marker-trait associations in potato. Molecular markers was acquired for relevant traits from a comprehensive literature study and tested on a diversity panel of 48 accessions of elite cultivars and breeding clones (manuscript 1). Even though it was necessary to design new primer sequences for most of the acquired markers (table 3, manuscript 1), this was possible to accomplish due to the application of the reference genome DM v4.03 (PGSC 2011) and the available genome sequences of the 18 MASPOT parents (Sverrisdóttir et al. 2017). From the converted markers presented in manuscript 1 and the markers from previous studies at Aalborg University in addition (manuscript 2) it has been possible to estimate the allele structure in 72 loci of interest. Hence, with the use of genome data obtained through converted markers in predictive modelling of phenotypic data (manuscript 2), this PhD study presents a successful utilization of existing molecular markers in predictive modelling.

The quality of historical data in predictive models

To the best of our knowledge this is the first time Historical data has been used for predictive modelling of potato phenotypes. Many studies base their analysis on populations that are thoroughly phenotyped at the onset of the study (kilde). Due to high expenses of phenotyping the population is normally limited to only a few cultivars. Utilizing historical data enables the use of larger populations without increasing the phenotyping cost (kilde). In this study we have used historical data from Danespo R&D, spanning the years from 1985-2014 in our predictive models. In the use of this data, we have encountered issues regarding different irregularities in the data set, such as varying number of recorded phenotypic values in the diversity panel dependent on trait. For the traits starch content and tuber shape phenotypic values were available for 40 and 39 of the 48 accessions in the diversity panel, while phenotypic values were only available for 16 out of 48 accessions for resistance against wart disease pathogen 6. Except for chip quality, phenotypic values for the traits that are tested by the breeding station (e.g. yield, starch content, tuber shape and late blight resistance) were of higher abundance, than for traits that are tested externally (e.g. resistance against nematodes). Chip quality testing is only performed routinely on cultivars intended for the processing industry, and phenotypic values of chip quality is therefore scarce in the diversity panel, which consist of cultivars from different market segments. Apart from the traits described in the manuscripts and additional results of this thesis, traits for resistance against different pathogens of *G. rostochiensis* and resistance against *G.*

pallida pathogen 3 were also examined for availability of historical data, but the availability was insufficient.

Distributions of phenotypic values also varied, as was shown in manuscript 2 and figure 1 of the additional results. It is expected that traits dependent on dominant resistance genes have groupings of phenotypic values into two distinct groups, and that was indeed the case for the three disease resistance traits covered in the additional results (figure 1C, D and E). However, in manuscript 2 it was shown that resistance against late blight had a more even distribution of phenotypic values of the diversity panel, though many dominant resistance genes have been associated with late blight resistance (Kilder). This could be because of the difference in years in which each cultivar was phenotyped. When there is a shift in the late blight pathogen population the inoculum used for infection of late blight resistance test fields will change in composition of pathogens. Hence, the phenotypic value of late blight resistance based on historical data will be corresponding to resistance against different pathogens dependent on which year the resistance was measured. Despite this, the predictive models for late blight resistance discussed in manuscript 2 yield similar results (0.6-0.85) to previous findings of prediction accuracy (0.7-0.8, Stich et al. 2018). Based on these findings, we have shown that it is indeed possible to obtain useful phenotypic values from historical data, though precautions must be taken regarding availability of phenotypic value for cultivars from different segments as well as potential difference in test criteria dependent on test year.

Pros and cons of predictive modelling

Prediction accuracy of the models in this study depended on the traits in question, which were partly due to difference in training population size between traits, because of varying availability of historic data. However, when the models for resistance against wart disease pathogen 1 and 6 are compared (pathogen 1: figure 3D and figure 6D, pathogen 6: figure 4 and figure 7 for SGP and RT models respectively), the RT model for pathogen 6 performed significantly better than the RT model for pathogen 1, while the opposite was true for the SGP models. Hence, the choice of predictive model approach is of great importance to the performance of the predictive models. This is most likely due to difference in model calculations and whether the trait is polygenetic or is affected by few dominant genes. The SGP models generally had lower RMSE values and were therefore more robust than the RT models both in results in manuscript 2 and in additional results. On the other hand, RT models had higher or similar coefficient of determination in 6 out of 9 traits examined in manuscript 2 (figure 4 and 5) and in the additional results (figure 2-7) and therefore had a higher prediction accuracy. From this, it is clear that different prediction modelling approaches have different strengths and

weaknesses. On the other hand, it might not be so clear which of the prediction modelling approaches is best suited for the trait in question. Fortunately, both approaches utilize genomic data as predictor variables and phenotypic data as response variables, so it is possible to follow both approaches without further experimental work and evaluate which approach is best suited for the trait in question.

In the context of prediction accuracy of the models, it is also worth considering, that the high values of coefficient of determination both in manuscript 2 and in the additional results, might be due to relatedness of the population, as was reported by Stich and Melchinger (2009). While the diversity panel was composed to be as diverse as possible based on estimates of inbreeding, all accessions are still from Danespo R/D breeding germplasm. Consequently, 13 cultivars and breeding clones recur in the ancestry of the accessions of the diversity panel, either as parent or grandparent. Furthermore, due to ancestry of many cultivars being unknown or of commercial confidentiality, 22 out of 48 accessions have one or more unknown grandparents. The estimated coefficient of inbreeding used to select accessions for the diversity panel is therefore expectedly underestimated. Sverrisdóttir et al. (2018) saw the largest drop in prediction correlations when models were trained with a population from a different geographic region than the test population. Derived from this, a solution for reduction of relatedness of the population could be composing the training population of cultivars from different breeding companies, but the difference in phenotype scoring method between companies becomes an obstacle for such a combined population. As there are no globally established standards for growth conditions and phenotypic testing of traits such as yield, starch content or chip quality, a combined population would need thorough, costly and time-consuming phenotypic testing before initiation of predictive modelling. While this would certainly result in models more widely applicable for different breeding companies, arguably, it is of higher importance for each breeding company to have a functional model at a lower cost, than to have a model optimized to include only genetic effects. The model would then have to be recalibrated after a certain timespan, when new cultivars have entered the breeding germplasm, to account for the change in relatedness.

An embedded condition in all predictive models in this thesis is the high correlation between markers resulting in collinearity. Increasing the resolution of allele structure by utilizing haplotype information as well as selecting locus markers in gene rich regions of the genome, results in increased linkage between haplotypes in close proximity. Furthermore, some markers are likely correlated because of chance co-inheritance due to relatedness and the small size of the diversity population. Correcting for correlation between markers is essential when

performing association studies to counter false positive results, since it becomes difficult to separate the effects of each marker when collinearity is present (James et al. 2013, D'hoop et al. 2014). Because of the collinearity of markers in this thesis (manuscript 2, supplementary) it is not possible to determine which of the selected locus markers are better associated with the trait in question, and in fact other locus markers might have resulted in models of the same prediction accuracy (James et al. 2013). The markers in this study was selected from a range of markers for different traits, and at a later stage, a selection of markers (35 out of 72) were analyzed on an offspring population. This meant a mix of markers for different traits were used to predict different traits, resulting in some traits being represented by a larger number of markers than others. In addition, because of the collinearity of markers, even a selection of markers resulting in high model prediction accuracy in the diversity panel was not guaranteed to produce the same results in an offspring population (manuscript 3). Because of the high collinearity of markers, a high risk of loss of linkage arises when the marker assay is reduced before employment to another population. This could be an explanation for the lower than expected coefficients of determination in manuscript 3.

Future perspectives

Regarding the implementation of molecular markers for in breeding context Slater et al. (2014) lists requirements for use of DNA-based markers to be effective. Among other conditions, the marker assay must be consistently reproducible, the procedure needs to be straight forward, the use must be cost-effective when compared to conventional procedures, and the associations between alleles and phenotypes should be applicable across different populations (Slater et al. 2014). The first and second conditions are met by our methodology, which exploits the standardized PCR amplifications and scripts. Regarding the cost-effectiveness, some phenotypic tests are readily suitable for replacement with predictive modelling despite the high prediction error of some models presented in this study. Resistance to nematodes is an example of such tests due to high variance of results and high cost of the testing. Expectedly, further testing of the locus markers on different and larger populations would result in models with lower prediction errors. With that, prediction of traits such as yield with higher complexity, should also become cost-effective compared to field screening procedures and in addition, further testing would make the models more applicable independent on population. One of the major obstacles in implementation of Genomic selection and prediction is the high start-up costs (Sverrisdóttir et al. 2017), which are lowered when using the methodology presented in this thesis. As the sequencing can be outsourced, the only required equipment is standard machinery, often already present in breeding station R&D departments.

References

ARRUDA, M.P., BROWN, P.J., LIPKA, A.E., KRILL, A.M., THURBER, C. and KOLB, F.L., 2015. Genomic Selection for Predicting Fusarium Head Blight Resistance in a Wheat Breeding Program. The Plant Genome, 8(3).

BARRELL, P.J., MEIYALAGHAN, S., JACOBS, J.M.E. and CONNER, A.J., 2013. Applications of biotechnology and genomics in potato improvement. Plant Biotechnology Journal, 11(8), pp. 907-920.

BOTSTEIN, D., WHITE, R.L., SKOLNICK, M. and DAVIS, R.W., 1980. Construction of a Genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms.

BREIMAN, 1984. Classification and Regression Trees. Routledge.

D'HOOP, B., PAULO, M., MANK, R., ECK, H. and EEUWIJK, F., 2008. Association mapping of quality traits in potato (Solanum tuberosum L.). Euphytica, 161(1), pp. 47-60.

DE JONG, W., LEISTER, D., GEBHARDT, C. and BAULCOMBE, D.C., 1997. A potato hypersensitive resistance gene against potato virus X maps to a resistance gene cluster on chromosome 5. TAG Theoretical and Applied Genetics, 95(1), pp. 246-252.

D'HOOP, B.B., KEIZER, L.C.P., PAULO, M.J., VISSER, R.G.F., EEUWIJK, V., F.A and ECK, V., H.J, 2014. Identification of agronomically important QTL in tetraploid potato cultivars using a marker-trait association analysis. Theoretical and Applied Genetics, 127(3), pp. 731-748.

ENCISO-RODRIGUEZ, F., DOUCHES, D., LOPEZ-CRUZ, M., COOMBS, J. and DE, L.C., 2018. Genomic Selection for Late Blight and Common Scab Resistance in Tetraploid Potato (Solanum tuberosum).

ESNAULT, F., PELLÉ, R., DANTEC, J., BÉRARD, A., LE PASLIER, M. and CHAUVIN, J., 2016. Development of a Potato Cultivar (Solanum tuberosum L.) Core Collection, a Valuable Tool to Prospect Genetic Variation for Novel Traits. Potato Research, 59(4), pp. 329-343.

FINKERS-TOMCZAK, A.M., DANAN, S., DIJK, V., T, BEYENE, A., BOUWMAN-SMITS, L., OVERMARS, H.A., ECK, V., H.J, GOVERSE, A., BAKKER, J. and BAKKER, E.H., 2009. A high-resolution map of the Grp1 locus on chromosome V of potato harbouring broad-spectrum resistance to the cyst nematode species Globodera pallida and Globodera rostochiensis. Theoretical and Applied Genetics, 119(1), pp. 165-173.

GEBHARDT, C., RITTER, E., DEBENER, T., SCHACHTSCHABEL, U., WALKEMEIER, B., UHRIG, H. and SALAMINI, F., 1989. RFLP analysis and linkage mapping in Solanum tuberosum. TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik, 78(1), pp. 65-75.

GROTH, J., SONG, Y., KELLERMANN, A. and SCHWARZFISCHER, A., 2013. Molecular characterisation of resistance against potato wart races 1, 2, 6 and 18 in a tetraploid population of potato (Solanum tuberosum subsp. tuberosum). Journal of Applied Genetics, 54(2), pp. 169-178.

HABYARIMANA, E., PARISI, B. and MANDOLINO, G., 2017. Genomic prediction for yields, processing and nutritional quality traits in cultivated potato (Solanum tuberosum L. Plant Breeding, 136(2), pp. 245-252.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition edn. New York, NY: Springer New York.

HESLOT, N., JANNINK, J. and SORRELLS, M.E., 2015. Perspectives for genomic selection applications and research in plants. 55(1), pp. 1.

HGSC, HUMAN GENOME SEQUENCING CONSORTIUM, 2001. Initial sequencing and analysis of the human genome. Nature, 409(6822), pp. 860.

HIRSCH, C., BUELL, C. and HIRSCH, C., 2016. A Toolbox of Potato Genetic and Genomic Resources. American Journal of Potato Research, 93(1), pp. 21-32.

HORTON, D.E., 1980. The potato as a food crop for the developing world. International Potato Center.

JACCOUD, D., PENG, K., FEINSTEIN, D. and KILIAN, A., 2001. Diversity Arrays: a solid state technology for sequence information independent genotyping. Nucleic Acids Research, 29(4), pp. 25.

JAMES, G., WITTEN, D., HASTIE, T., TIBSHIRANI, R. and SPRINGERLINK (ONLINE SERVICE), 2013. An Introduction to Statistical Learning : with Applications in R / by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. Elektronisk udgave edn. New York, NY : Springer New York.

JANSEN VAN RENSBURG, W. S and DUBERY, I.A., 2001. Development of a sequence characterized amplified region (SCAR) marker for the identification of the potato cultivars Astrid and Mnandi. South African Journal of Plant and Soil, 18(4), pp. 154-158.

KAUL, S. and K., 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature, 408(6814), pp. 796.

KIT, S., 1961. Equilibrium sedimentation in density gradients of DNA preparations from animal tissues.

KONIECZNY, A. and AUSUBEL, F.M., 1993. A procedure for mapping Arabidopsis mutations using co-dominant ecotype-specific PCR-based markers. Plant Journal.

KREIKE, C.M., DE KONING, J. R. A., VINKE, J.H., VAN OOIJEN, J.W. and STIEKEMA, W.J., 1994. Quantitatively-inherited resistance to Globodera pallida is dominated by one major locus in Solanum spegazzinii. TAG Theoretical and applied genetics, 88(6-7), pp. 764-769.

LI, L., PAULO, M.J., EEUWIJK, V., F.A and GEBHARDT, C., 2010. Statistical epistasis between candidate gene alleles for complex tuber traits in an association mapping population of tetraploid potato. Theoretical and Applied Genetics, 121(7), pp. 1303-1310.

MEUWISSEN, T.H., HAYES, B.J. and GODDARD, M.E., 2001. Prediction of total genetic value using genome- wide dense marker maps. Genetics, 157(4), pp. 1819-1829.

MILCZAREK, D., FLIS, B. and PRZETAKIEWICZ, A., 2011. Suitability of Molecular Markers for Selection of Potatoes Resistant to Globodera spp. American Journal of Potato Research, 88(3), pp. 245-255.

MOLONEY, C., GRIFFIN, D., JONES, P., BRYAN, G., MCLEAN, K., BRADSHAW, J. and MILBOURNE, D., 2010. Development of diagnostic markers for use in breeding potatoes resistant to Globodera pallida pathotype Pa2/3 using germplasm derived from Solanum tuberosum ssp. andigena CPC 2802. Theoretical and Applied Genetics, 120(3), pp. 679-689.

MORRELL, P.L., BUCKLER, E.S. and ROSS-IBARRA, J., 2012. Crop genomics: advances and applications. Nature Reviews Genetics, 13(2), pp. 85-96.

MUKAKA, M.M., 2012. Statistics corner: A guide to appropriate use of correlation coefficient in medical research.

PARAN, I. and MICHELMORE, R.W., 1993. Development of reliable PCR-based markers linked to downy mildew resistance genes in lettuce. TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik, 85(8), pp. 985.

POCZAI, P., VARGA, I., LAOS, M., CSEH, A., BELL, N., VALKONEN, J.P. and HYVÖNEN, J., 2013. Advances in plant gene-targeted and functional markers: a review. Plant methods, 9(1), pp. 6.

RAMAKRISHNAN, A.P., RITLAND, C.E., BLAS SEVILLANO, R.H. and RISEMAN, A., 2015. Review of Potato Molecular Markers to Enhance Trait Selection. American Journal of Potato Research, 92(4), pp. 455-472.

REID, A., HOF, L., FELIX, G., RUCKER, B., TAMS, S., MILCZYNSKA, E., ESSELINK, G., UENK-STUNNENBERG, G.E., VOSMAN, B. and WEITZ, A., 2011. Construction of an integrated microsatellite and key morphological characteristic database of potato varieties on the EU common catalogue. Euphytica, 182(2), pp. 239-249.

REJWAN, C., COLLINS, N.C., BRUNNER, L.J., SHUTER, B.J. and RIDGWAY, M.S., 1999. Tree regression analysis on the nesting habitat of smallmouth bass. Ecology, 80, pp. 341-348.

ROUPPE VAN DER VOORT, J. N. A. M, ECK, V., H.J, DRAAISTRA, J., ZANDVOORT, V., P.M, JACOBSEN, E. and BAKKER, J., 1998. An online catalogue of AFLP markers covering the potato genome. Molecular Breeding, 4(1), pp. 73-77.

SLATER, A.T., COGAN, N.O.I., FORSTER, J.W., HAYES, B.J. and DAETWYLER, H.D., 2016. Improving Genetic Gain with Genomic Selection in Autotetraploid Potato. Plant Genome, 9(3), pp. 1-15.

SLATER, A., COGAN, N., HAYES, B., SCHULTZ, L., DALE, M., BRYAN, G. and FORSTER, J., 2014. Improving breeding efficiency in potato using molecular and quantitative genetics. Theoretical and Applied Genetics, 127(11), pp. 2279-2292.

ŚLIWKA, J., JAKUCZUN, H., CHMIELARZ, M., HARA-SKRZYPIEC, A., TOMCZYŃSKA, I., KILIAN, A. and ZIMNOCH-GUZOWSKA, E., 2012. A resistance gene against potato late blight originating from Solanum × michoacanum maps to potato chromosome VII. Theoretical and Applied Genetics, 124(2), pp. 397-406.

STICH, B. and MELCHINGER, A.E., 2009. Comparison of mixed-model approaches for association mapping in rapeseed, potato, sugar beet, maize, and Arabidopsis. BMC genomics, 10(1), pp. 94.

STICH, B., URBANY, C., HOFFMANN, P., GEBHARDT, C. and LÉON, J., 2013. Population structure and linkage disequilibrium in diploid and tetraploid potato revealed by genome-wide high-density genotyping using the SolCAP SNP array. Plant Breeding, 132(6), pp. 718-724.

STICH, B. and VAN INGHELANDT, D., 2018. Prospects and Potential Uses of Genomic Prediction of Key Performance Traits in Tetraploid Potato. Frontiers in Plant Science, 9.

SULLI, M., MANDOLINO, G., STURARO, M., ONOFRI, C., DIRETTO, G., PARISI, B. and GIULIANO, G., 2017. Molecular and biochemical characterization of a potato collection with contrasting tuber carotenoid content. PLoS One, 12(9), pp. e0184143.

SVERRISDÓTTIR, E., BYRNE, S., SUNDMARK, E., JOHNSEN, H., KIRK, H., ASP, T., JANSS, L. and NIELSEN, K., 2017. Genomic prediction of starch content and chipping quality in tetraploid potato using genotyping-by-sequencing. Theoretical and Applied Genetics, 130(10), pp. 2091-2108.

SVERRISDÓTTIR, E., SUNDMARK, E.H.R., JOHNSEN, H.Ø, KIRK, H.G., ASP, T., JANSS, L., BRYAN, G. and NIELSEN, K.L., 2018. The Value of Expanding the Training Population to Improve Genomic Selection Models in Tetraploid Potato. Frontiers in plant science, 9, pp. 1118.

TIWARI, J.K., SIDDAPPA, S., SINGH, B.P., KAUSHIK, S.K., CHAKRABARTI, S.K., BHARDWAJ, V., CHANDEL, P. and WEHLING, P., 2013. Molecular markers for late blight resistance breeding of potato: an update. Plant Breeding, 132(3), pp. 237-245.

TOMCZYŃSKA, I., STEFAŃCZYK, E., CHMIELARZ, M., KARASIEWICZ, B., KAMIŃSKI, P., JONES, J., LEES, A. and ŚLIWKA, J., 2014. A locus conferring effective late blight resistance in potato cultivar Sárpo Mira maps to chromosome XI. Theoretical and Applied Genetics, 127(3), pp. 647-657.

VALIN, H., SANDS, R.D., VAN, D.M., NELSON, G.C., AHAMMAD, H., BLANC, E., BODIRSKY, B., FUJIMORI, S., HASEGAWA, T., HAVLIK, P., HEYHOE, E., KYLE, P., MASON-D' CROZ, D., PALTSEV, S., ROLINSKI, S., TABEAU, A., VAN MEIJL, H., VON LAMPE, M. and WILLENBOCKEL, D., 2014. The future of food demand: understanding differences in global economic models. Agricultural Economics, 45(1), pp. 51-67.

VAN-ECK, H.J., JACOBS, J., STAM, P., TON, J., STIEKEMA, W.J. and JACOBSEN, E., 1994. Multiple Alleles for Tuber Shape in Diploid Potato Detected by Qualitative and Quantitative Genetic Analysis Using RFLPs. Genetics, 137(1), pp. 303-309.

VOS, P., HOGERS, R., BLEEKER, M., REIJANS, M., LEE, T.V.D., HORNES, M., FRITERS, A., POT, J., PALEMAN, J., KUIPER, M. and ZABEAU, M., 1995. AFLP: a new technique for DNA fingerprinting. Nucleic Acids Research, 23(21), pp. 4407-4414.

WANG, D.G., FAN, J., SIAO, C., BERNO, A., YOUNG, P., SAPOLSKY, R., GHANDOUR, G., PERKINS, N., WINCHESTER, E., SPENCER, J., KRUGLUAK, L., STEIN, L., HSIE, L., TOPALOGLOU, T., HUBBELL, E., ROBINSON, E., MITTMANN, M., MORRIS, M.S., SHEN, N., KILBURN, D., RIOUX, J., NUSBAUM, C., ROZEN, S., HUDSON, T.J., LIPSHUTZ, R., CHEE, M. and LANDER, E.S., May 15, 1998-last update, Large-Scale Identification, Mapping, and Genotyping of Single-Nucleotide Polymorphisms in the Human Genome on JSTOR. Available: https://www-jstor-org.zorac.aub.aau.dk/stable/2895444 [Sep 3, 2018].

WILLIAMS, J., 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic acids research, 18(22), pp. 6531-6535.

XU, X., PAN, S.K., CHENG, S.F., ZHANG, B., BACHEM, C.W.B., BOER, D., J.M, BORM, T.J.A., KLOOSTERMAN, B.A., ECK, V., H.J, DATEMA, E., GOVERSE, A., HAM, V., R.C.H.J and VISSER, R.G.F., 2011. Genome sequence and analysis of the tuber crop potato. Nature, 475(7355), pp. 189-195.

XU, Y. and CROUCH, J.H., 2008. Marker-Assisted Selection in Plant Breeding: From Publications to Practice. Crop Science, 48(2), pp. 391.

ISSN (online): 2446-1636
ISBN (online): 978-87-7210-436-2

AALBORG UNIVERSITY PRESS